proteins. The sequence-specific terms were derived as statistical potentials with a rather careful selection of the reference state.[20,25,26] When several statistical potentials are combined in a relatively complex reduced model, an *a priori* derivation of the relative scaling factors becomes difficult. Some double counting of particular physical interactions may occur. Thus, these scaling factors have to be adjusted to reproduce a reasonable balance between the short- and long-range interactions. A proper balance should lead to a low secondary structure content in the denatured state and a well-packed and ordered collapsed state. The collapse transition should be as abrupt as possible, mimicking an all-or-none folding transition. This has been achieved in the present model with the given scaling of particular interactions. Folding experiments for several proteins of various structural classes were performed with no short- or long-range constraints. The force field described above fails to produce a unique folded state, except for very simple folding motifs. For more complex motifs, the folded states always had a secondary structure very close to the native, with good packing of the hydrophobic core; however, the arrangement of the secondary structure elements (connection of helices, order of β-strands in sheets, *etc.*) almost always had topological errors. As designed, the model with its force field is very efficient at generating protein-like compact conformations. The model is not sensitive to the particular scaling of the various interactions within a broad range around the set used in this work. For example, removal of all generic terms also led to collapsed structures (although at lower temperatures) with good overall fidelity of the secondary structure, but the geometrical accuracy of the secondary structure and packing pattern was more irregular. A detailed discussion of the interplay between the generic and sequence-specific short-range potentials is reported elsewhere.[27] When the proposed force field is supplemented by one or more structural constraints, a proper fold should be easily selected.

Since a Cα-based MONSTTER model has been reported as being successful in reproducing quite complex aspects of protein dynamics and

thermodynamics,[6,15,16,23,28-36] without being bound to any particular theory, it is believed that the present force field approximately reproduces the main features of globular proteins. However, it does so in a different geometrical context, namely using pseudoatoms representing side chain centers of mass. Moreover, the instant invention is based on a less complex representation and simpler definition of the force field, and is more computationally more efficient than Cα-based models such as MONSSTER. As a result, three-dimensional structures for larger proteins can be simulated.

Physical Basis of the Model Interaction Scheme

The instant invention allows realistic three-dimensional protein structures (as seen on the level of an entire fold) to be produced from an extremely simplified representation of the protein conformational space. Here, only the side chains (in one embodiment represented by their respective centers of mass) are explicitly modeled. The use of a single interaction unit per residue is computationally very efficient. Moreover, side chains were used, as opposed to, for example, alpha-carbons, because the specific interactions between, or functions of, proteins involve side chains, while main chain (*i.e.*, peptide backbone) interactions are much less dependent on amino acid sequence. Due to this very simple representation and requested specificity, several features have to be built into the model force field. First, the assumed protein representation, with a single center of interaction per amino acid residue side chain, allows too much conformational freedom. This is because there is no explicit backbone connectivity in the model chains. However, in real proteins, the backbone connectivity and conformational stiffness control, to some extent, the distances between the centers of mass of the side groups near each other along the polypeptide chain. The backbone effect is moderated by the side chains' internal degrees of freedom. It is reasonable to assume that for a short polypeptide fragment, the local geometry of the side chain centers of mass is mostly dictated by short-range interactions with a somewhat lesser effect from long-range

(tertiary) interactions. The correct, protein-like distance geometry of the side chain centers of mass implies a correct, protein-like geometry of the main chain. This provides a conceptual background for the sequence-specific short-range potential of mean force (discussed previously and defined in equation 1, above). This potential drives the system towards a local geometry (characterized by distances between side chains) that is characteristic of locally similar sequences.

At first glance, it may appear that such defined sequence-specific secondary propensities are sufficient for modeling protein like local geometry. This is not the case for several reasons. First, the discussed statistical potentials are not very accurate due to the limited size of the database of known protein structures. However, more importantly, the assumed simplified representation of the polypeptide chains exhibits excessive flexibility. With respect to the assumed model of excluded volume, a substantial fraction of the model chain conformations that are otherwise allowed are conformations that cannot possibly occur in any protein or even in other polymers. It is not a good strategy to make the sequence-specific interactions so strong that the non-physical geometries would be practically prohibited. This would lead to dynamic frustration of the model system due to very frequent trapping in the local conformational energy minima; thus, providing a generic bias towards protein-like geometry is computationally more efficient. Then, much less is required from the sequence-specific part of the potential (selection within the protein-like part of conformational space instead of selection within a much larger conformational space of a freely joined polymer chain). Moreover, a properly defined generic potential can "interpolate" protein-like conformations for those fragments of a given polypeptide chain where the information content of the sequence-specific potential is low, (due to lack of examples in the database or balanced contradictory examples). As discussed above (*see* equations 3 and 4), sequence-independent potentials exactly play such a role. The first such term provides a bias towards the protein-like stiffness of the model chain by an energetic preference for either expanded zigzag or helical conformations. The second term